CTGANおよびTVAEを活用したアクティビティデータ生成手法の構築

石嶋　悠嗣[1]・柳沼 秀樹[2]・寺部　慎太郎[3]・海野　遥香[4]・鈴木　雄[5]

[1]学生非会員　東京理科大学大学院　理工学研究科土木工学専攻（〒 278-8510 千葉県野田市山崎 2641）
E-mail: 7621502@ed.tus.ac.jp
[2]正会員　東京理科大学准教授　理工学部土木工学科（〒 278-8510 千葉県野田市山崎 2641）
E-mail: yaginuma@rs.tus.ac.jp
[3]正会員　東京理科大学教授　理工学部土木工学科（〒 278-8510 千葉県野田市山崎 2641）
E-mail: terabe@rs.tus.ac.jp
[4]正会員　東京理科大学助教　理工学部土木工学科（〒 278-8510 千葉県野田市山崎 2641）
E-mail: unoharuka@rs.tus.ac.jp
[5]正会員　東京理科大学助教　理工学部土木工学科（〒 278-8510 千葉県野田市山崎 2641）
E-mail: yusuzuki@rs.tus.ac.jp

近年，交通行動分析においてもニューラルネットワーク等の機械学習理論を援用した研究事例が多く見られる．これらの手法を用いる際には，膨大な学習データが必要であり，かつデータの質が担保されていることが望ましい．しかしながら，個人の活動・交通行動データの取得コストは未だに高く，十分なデータを確保することが難しい状況にある．本研究では，GAN(Generative Adversarial Network) を用いた PT データの拡張手法を提案し，首都圏をケーススタディとしたデータセットの生成を試みる．具体的には，LSTM(Long Short Term Memory) を組み込んだ CTGAN(Conditional Tabular GAN) や Autoencoder を活用した TVAE(Triplet-based VAE) を適用することで前後のアクティビティ情報を保持したデータ生成が可能になった．これにより小サンプルでのモデル学習やデータ収集の効率化が期待される．

Key Words: *Neural Network, GAN(Generative Adversarial Network), Data generation*

1. はじめに

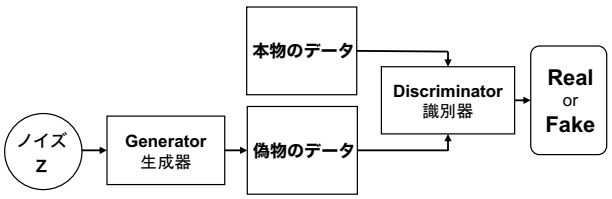
近年，交通行動分析においてもニューラルネットワーク等の機械学習理論を援用した研究事例が多く見られる．特に交通量計測や災害時の異常検出においては，分析手法や出力結果に対する解釈性が不要であること，長期的かつ継続的な観測により十分量のデータセットが確保されている点などから機械学習の導入が進んできている．一方，交通行動分析においては機械学習の導入は進んでいない．主に理由は 2 つあると考えられる．1 つは交通行動予測には「解釈性」が重要視されるため，人間が理解できるモデルでなければ政策への適用が困難であるという点だ．機械学習モデルの予測性能は高性能だが解釈性は乏しいため，モデルから選択行動の要因を示すことが難解である．そのため実務への適用が難しく政策への導入が遅れている可能性が考えられる．2 つ目は，十分量のデータセットを確保できない点である．機械学習を用いる際にはデータの量や質が担保されていることが望ましく，学習データが十分に確保できればモデルの汎化性能の向上や過学習の抑制にも繋がる．しかしながら，個人の活動・交通行動データの取得コストは未だに高い．例えば，東京 PT データでは十分量のデータセットは確保されているものの，調査から集計には長い期間を要しており，データ収集の効率性は低い状態にある．そこで本研究では，GAN(Generative Adversarial Network) 及び VAE(Variational Autoencoder) を用いた PT データ生成手法を提案し，首都圏をケーススタディとしたデータセット生成を試みる．具体的には LSTM(Long Short Term Memory) を用いた CTGAN と Autoencoder を活用した TVAE(Triplet-based VAE) を適用し，前後のアクティビティ情報を加味したデータ生成を行う．さらに生成データが機械学習のデータセットとして適用可能であるかを検証するために，生成データと本物データを用いて予測モデルの構築を行い精度の検証を行った．2 つの生成モデルの比較及び検討を行い，データ収集の効率化と小サンプルでの安定学習を目指す．

2. 本研究で用いるデータ生成手法

(1) CTGAN

a) モデルの概要

画像生成の分野において，膨大な学習データを用いることによって実画像と見分けがつかないような偽画



図–1 GAN(Generative Adbersarial Network)

像を生成する敵対的生成ネットワーク (GAN) が注目されている．**図–1** に示すように GAN のネットワーク構造は，Generator（生成器）と Discriminator（識別器）の 2 つで構成されており，この 2 つのネットワークを競い合わせることで精度を高めるモデル構造である．生成器は識別器が本物と誤判定するような偽データを生成するようにパラメータを更新し，反対に識別器は本物データか偽物データかを正確に判断できるようにパラメータを更新することによって，生成器が実際に存在しない偽のデータを生成することが可能になる．生成器がデータセットと同様の確率分布を上手く再現し，その分布に従ってサンプルすることでデータを生成している．また GAN は教師なし学習であるため，従来の深層学習のようにデータラベルを必要としない学習が特徴であるが，ラベリングがないことによって学習が不安定になるため学習を安定させるような工夫が必要である．

本研究では，この GAN を用いたテーブルデータ生成手法である CTGAN によって東京 PT データの拡張を行う．そもそも GAN は画像データを生成するために開発された手法であったが，非画像データに対する研究も進んできている．特に医療分野においては，個人情報保護の観点や投与データを十分に確保できないなどの問題点が挙げられており，これらの問題を解決するために GAN によるテーブルデータ生成手法の開発が進んできた[1),2),3),4),5),6)]．また，テーブルデータには連続値，離散値，日時，カテゴリー変数など様々なデータが内包されており，どのデータタイプに対しても適応することが求められている．そして CTGAN はそれらの課題を考慮してデータ生成を行うことが可能である．

b) モデルの構造

前述したように，本研究では CTGAN を用いてデータ生成を行う．CTGAN の先行研究として，Lei Xu らの TGAN[7)] がある．これは生成器に LSTM(Long Short Term Memory)，識別器に MLP（Multi-Layer Perceptron）を適用したモデルである．また，連続値に対しては GMM(Gaussian Mixture Model) を用いて各カラム複雑な分布を特定することによって正確に正規化を行い，離散値に対してはランダムなノイズを加えることによって滑らかな分布にするような工夫をしている．これにより TGAN は他のテーブルデータ生成モデルよりも高品質なデータを生成することが可能になった．しかし，TGAN はマルチモーダル分布での正規化が不十分である点や離散カラムに対しての不均衡性などの問題があったため，その課題に対処するためのモデルとして CTGAN が開発された．

まず CTGAN では，1 つ目のマルチモーダル分布に対するカラムでの正規化方法の問題に対しては VGM(Variational Gaussian Mixture model) を用いることで対処している．手順としては，まず VGM によりマルチモーダル分布を複数の正規分布に分解し，それらの分布を元に連続量が各分布にどのくらいの確率で選択されるかを計算し，その確率の元で正規化する分布を適宜選んでいく．そして正規化された値と選ばれた分布を one-hot 化した値が変換された連続量となる．これらの処理を行なった連続値のカラムと one-hot 化した離散値のカラムを結合したものがデータセットとして表現される．

2 つ目の離散値のカラムに対する不均衡性に関しては，条件ベクトルと Training-by-Sampling という方法を用いることで対処している．多くの場合，与えられたデータセットのカテゴリー変数は偏りが生じていることが多く，その偏りが学習時にも影響していると考えられる．例えば，ランダムにサンプリングして学習した場合，出現頻度が低いマイナーなカテゴリー変数は十分に抽出されず，正確に学習できない可能性が考えられる．これらを対処するために離散値を one-hot 化した条件ベクトルを学習時の変数として追加し，この条件ベクトルを元に均等にサンプリング (Training-by-sampling) することによって不均衡性を避けている．

今回の分析では，データ生成ツール SDV[8),9)] を用いて CTGAN の学習を行った．生成器と識別器のネットワークは共に (256,256) であり，バッチサイズは 500，epochs 数は 300 回で学習を行った．

(2) VAE

a) モデルの概要

VAE は Kingma[10)] らによって提案された生成モデルであり，入力データを低次元の潜在変数に圧縮する Encoder と，潜在変数から入力データを復元して出力する Decoder で構成される．入力データと生成データの誤差が小さくなるように VAE を学習させることで，教師データを用いずにデータ生成が可能になる．また VAE は，再パラメータ化トリックを導入することにより，SGD のみで潜在変数付き Encoder-Decoder を実現できている点が VAE の工夫点である．従って，学習時の計算効率性もよく，他のニューラルネットワークと同様の仕

組みの SGD のみで学習が可能である．

b) モデルの構造

VAE のアルゴリズムを説明する．VAE では確率変数 x と潜在変数 z のデータ生成過程を，確率密度分布のモデルパラメータを用いて $z \sim p(z), x \sim p_\theta(x|z)$ のように定める．そして真の事後分布 $p_\theta(x|z)$ を近似した分布 $q_\theta(z|x)$（ϕ はモデルパラメータ）を用いて，周辺尤度の下界 $\mathcal{L}(\theta, \phi; x)$ が以下のように求まる．

$$\log p_\theta(x) = D_{KL}(q_\phi(z|x)||p_\theta(z|x)) + \mathcal{L}(\theta, \phi; x) \quad (1)$$

$$\begin{aligned} &\geq \mathcal{L}(\theta, \phi; x) \\ &= -D_{KL}(q_\phi(z|x)||p_\theta(z)) \\ &\quad + E_{q_\phi(z|x)}[\log p_\theta(x|z)] \end{aligned} \quad (2)$$

そして訓練では，下界 $\mathcal{L}(\theta, \phi; x)$ が最大となるようにパラメータ θ, ϕ について最適化を行う．$q_\theta(z|x)$ がエンコーダー，$p_\theta(x|z)$ がデコーダーとなる．

そして本研究で用いる TVAE の入力データは CTGAN と同様に，VGM で正規化された連続量と one-hot 化した離散値がモデルのインプットとなる．TVAE の学習も CTGAN と同様に，SDV を用いて分析を行った．エンコーダーとデコーダーのネットワークは共に (512,512) とし潜在変数の次元は 128，バッチサイズは 500,epochs 数は 300 回に固定し学習をおこなった．

(3) ニューラルネットワークによる予測

本研究では，生成データの有用性を検証するために，多層ニューラルネットワーク (MLP) による予測モデルを構築し，生成データが本物データ同様の予測が可能であるかを分析した．そしてデータセットは活動パターンと交通手段の 2 つを用意し，各データに対して MLP を学習させた．活動パターンと交通手段での MLP の中間層は (100,100,50) と固定し，出力層のニューロン数だけを変更した．そして，過学習を避けるために全結合層の後にバッチ正規化を加えた．また，検証用のデータとしてランダムに 1000 サンプル抽出しておき，そのデータセットに対して予測をすることで精度の検証を行っている．

3. アクティビティデータ生成手法の構築

(1) データの概要

本研究では，H30 東京パーソントリップ調査データ（以下，東京 PT データ）を用いた．東京 PT データは個人の平日 1 日のトリップデータを収集している．東京を中心とする半径約 80km 圏域を対象とした調査であり，平成 30 年の 9 月から 11 月に行われた．本研究ではこのデータを用いてデータ生成によるモデルの精度検証を行う．

今回はデータ生成の有効性を検証するために，東京 PT データから交通手段選択モデル用と活動パターン選択モデル用の 2 つのデータセットを用意した．データセットに使用した特徴量と目的変数は**表–1**, **表–2** に示す．この 2 つのデータセットに対して CTGAN と TVAE を適用する事でデータを生成し，新たなデータセットを確保することが目的である．また，データ生成をする際にはいくつかの条件が必要であるが，これらは SDV 内のモデルを使用すれば様々な制約条件を与えることで表現が可能である．例えば，到着時刻は出発時刻よりも大きいこと，到着時刻と出発時刻の差が移動時間になること，ある閾値よりも小さな値を生成しなければいけないなどデータセット内では複数の条件が考えられる．これらは SDV 内で自由に設定することが可能で，本データでは上記のように出発時刻，到着時刻，移動時間の整合が取れるような制約と，移動時間に関しては外れ値を生成させないようにするため，120 分以下のトリップを生成するように条件を与えて生成を行った．

表–1 活動パターンの特徴量と目的変数

特徴量	年齢，性別，移動目的， 収入，職業，移動時間， 滞在時間，出発施設，到着施設， 出発ゾーン，到着ゾーン， 出発時刻，到着時刻，代表交通手段
目的変数	活動パターン（全 27 選択肢）

表–2 交通手段の特徴量と目的変数

特徴量	年齢，性別，移動目的，収入， 職業，移動時間，滞在時間， 出発施設，到着施設，出発ゾーン， 到着ゾーン，出発時刻，到着時刻， トリップ回数，各交通手段の移動時間 各交通手段の費用，選択肢集合
目的変数	代表交通手段（全 5 選択肢）

(2) 生成データの評価

生成モデルの評価には KL ダイバージェンスと最尤推定量を用いる．KL ダイバージェンスは 2 つの確率分布の類似度を測定する尺度であり，0 に近いほど 2 つの分布が類似していると定義される．そして，生成モデルは連続量の分布を複数の正規分布が混合していると仮定しているため，正規分布ではなく GMM による KL ダイバージェンスを計算する必要がある．計算式は以

下のように表現される．

$$f(x) = \sum_a \pi_a N(x; \mu_a, \sigma_a) \tag{3}$$

$$g(x) = \sum_b \pi_b N(x; \mu_b, \sigma_b) \tag{4}$$

$$D_{MC}(f||g) = \frac{1}{n}\sum_i^n \log(\frac{f(x_i)}{g(x_i)}) \tag{5}$$

2 つの正規分布 $f(x)$，$g(x)$ の距離を計算したものが $D_{MC}(f||g)$ となる．まず，混合比 π_a に従って正規分布を 1 つ選び，選ばれた k 番目の正規分布 $N(;\mu_k, \sigma_k)$ から乱数 x_i を 1 つサンプリングする．そして $f(x_i)$,$g(x_i)$ を計算し，これを N 回繰り返して平均を取ることで求められる．本研究では，連続値のみの評価に用いた．

また，尤度推定量にも同様に GMM を用い，対数尤度関数 (6) を計算する．

$$\begin{aligned}\log L(X|\pi, \mu, \Sigma) &= \log\{\prod_{j=1}^N \sum_{k=1}^n \pi_k N(x_j|\mu_k, \Sigma_k)\} \\ &= \sum_{j=1}^N \log\{\sum_{k=1}^n \pi_k N(x_j|\mu_k, \Sigma_k)\}\end{aligned} \tag{6}$$

μ_k は k 番目の正規分布における $1 \times m$ の平均ベクトル，Σ_k は k 番目の正規分布における $m \times m$ の分散共分散行列，π_k は混合係数である．

本研究では，この 2 つの評価指標を用いて生成データの有効性を検証した．計算結果を**表–3** に示す．元のデータセットのうち 10000 サンプルを抽出し，それらに対して各指標を計算した．また，どちらの指標に関しても各カラムの結果を平均したものが最終的な値となっている．まず，KL ダイバージェンスに関しては，どちらのデータセットに関しても CTGAN の方が小さくなっており，TVAE よりも各連続量の分布の当てはまりが優れていることが確認された．また尤度に関しては，両データセットで TVAE の方が優れている結果となった．しかし，どちらも僅かな差であるため大きな影響はないと考えられる．

表–3 KL ダイバージェンスと尤度

	生成モデル	KL	対数尤度
活動パターン	CTGAN	0.879	-49.8
	TVAE	0.909	-47.6
交通手段	CTGAN	0.812	-87.3
	TVAE	0.907	-83.5

4. 提案手法の精度検証

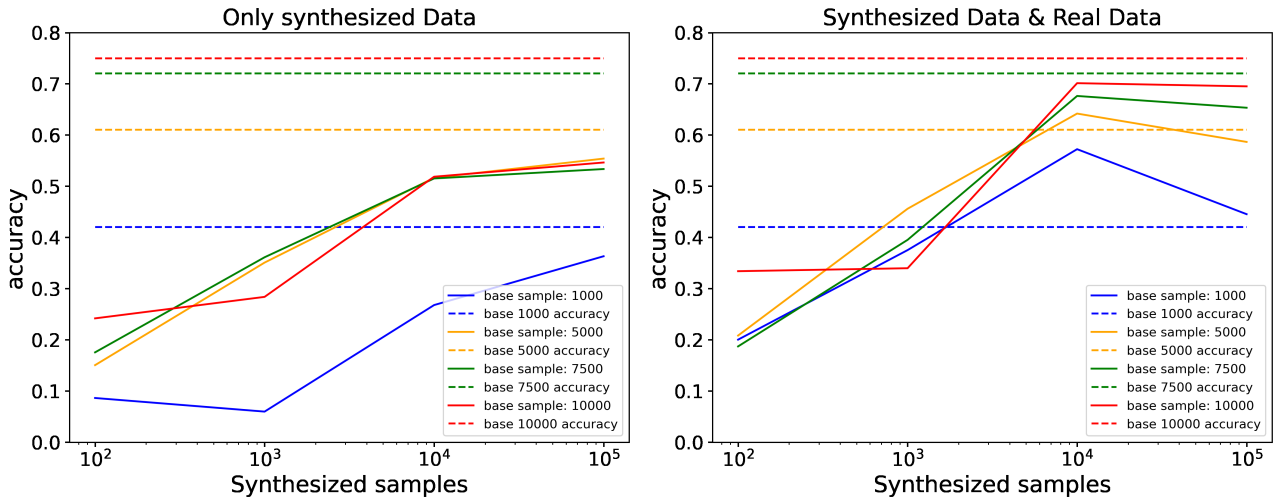
(1) CTGAN を活用した生成手法

CTGAN で生成したデータの検証結果を**図–2**, **図–3** に示す．ベースサンプルは分析者が確保できているデータ数を想定しており，これらの数を変更することでどれくらいのデータ数がデータ生成に有効であるかを検証した．またベースラインはベースサンプルでの精度を示しており，ベースサンプルが増えるほど精度が高くなっている．図-2，図-3 の左図は活動パターンと交通手段において生成データのみで推定を行った結果である．どのサンプル数においてもベースラインの精度を超えておらず，生成データの質は低い結果となった．また，生成データと実データを結合して推定した右図の結果では，少数サンプルでは精度向上が見込めるものの，それ以上のサンプルにおいては同等もしくはそれ以下の精度となっている．上手くデータを生成できていないため，本物データと結合しても精度向上が見込めない結果となった．

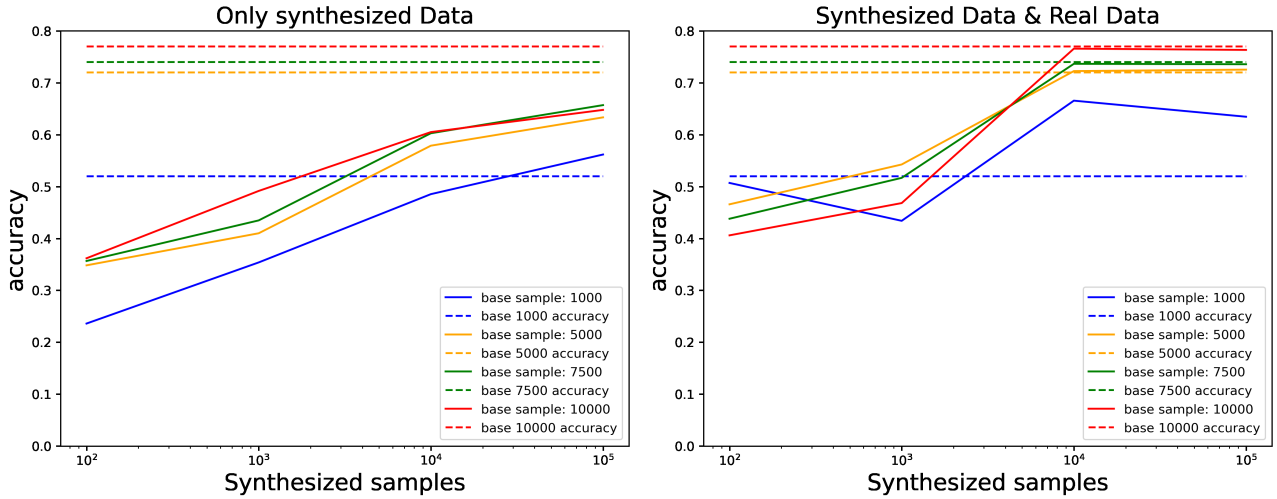
(2) TVAE を活用した生成手法

TVAE で生成したデータの検証結果を**図–4**, **図–5** に示す．図-4，図-5 の左図は CTGAN と同様に生成データのみで推定を行った結果である．どのベースサンプルでも生成するデータ数が多くなれば精度も向上しており，特にベースサンプルが 1000 個の時には的中率が大幅に向上し，データ生成の有効性が確認された．また，それ以上のデータ数に関しては，1000 サンプルほどの精度向上は確認されなかったが，ベースラインの精度を僅かに上回る，もしくは元のデータセットと同等の精度を得ることが確認された．よって，5000 サンプル以上では，予測モデルの学習に必要なデータ数は確保できており，データ生成による拡張は精度や汎化性能向上に大きく影響しないことが示唆された．一方，生成データと実データの結合データに関しては，生成データのみの場合と同様に，1000 サンプルでは精度の大幅な向上が確認され，それ以外のサンプル数に関しては，僅かではあるがベースラインの精度を上回っており，汎化性能向上が確認された．CTGAN に比べ生成データの学習がうまく行っているため，実データと合わせることでより汎用性の高い学習ができていると考えられる．

以上の結果から，PT データに関しては TVAE の方が優れていることが確認された．本研究において，生成データの評価としては尤度と KL ダイバージェンスを適用したが，これらの指標はあくまで元データと生成データの各カラムに対して計算し，それらを平均した



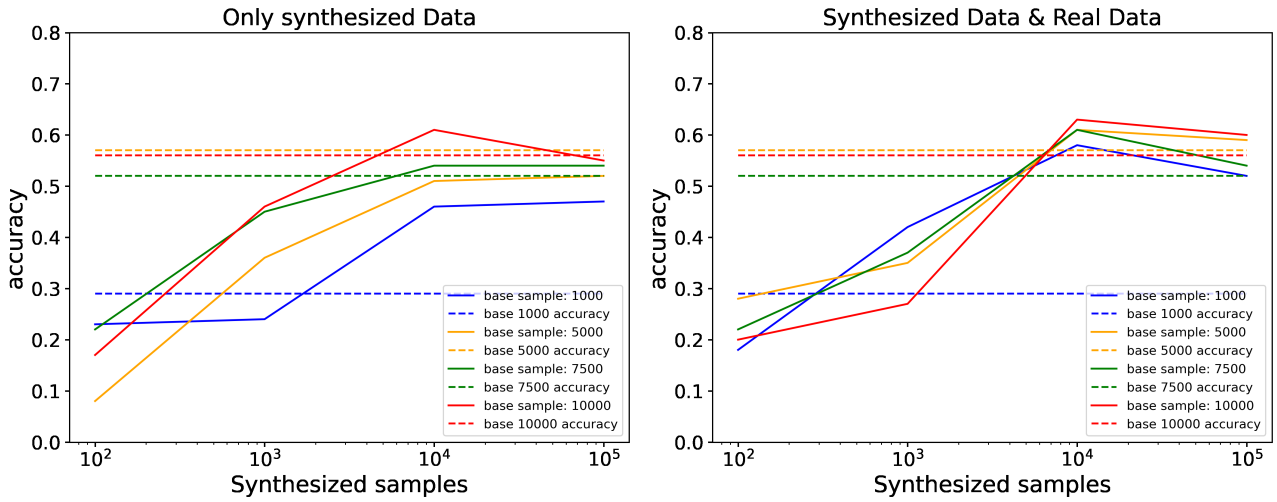
図–2 活動パターンの精度 (CTGAN)



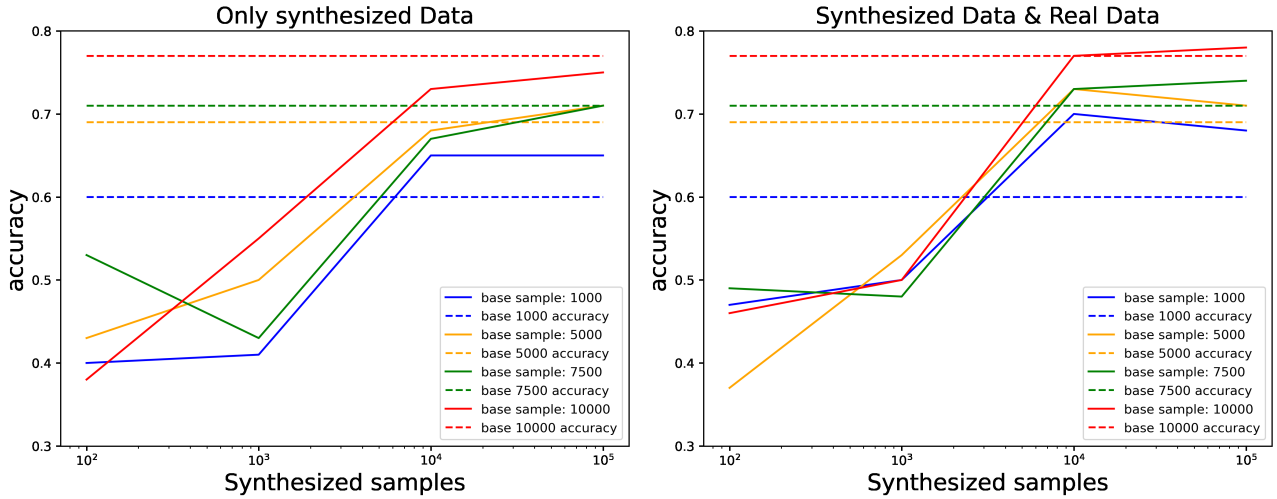
図–3 交通手段の精度 (CTGAN)

ものであるため，カラム単体の尤度や分布の類似度しか計算できていない．そのため，テーブルデータ全体の評価はできておらず，今回の指標だけでは生成データセットと本物のデータセットの類似度を判断することは困難であった．生成データ全体の有用性を考慮するためにも，生成されたデータを用いて元データとの精度比較を行った検証方法が有効であり，今回の分析で生成データ全体の有効性を示すことができた．

以上のことから，精度向上を図る際には少数データに対してデータ拡張を行うことは有効であり，それ以上のサンプルになると大きな精度向上は見込めず僅かな汎用性向上にとどまることが確認された．これらを踏まえると，データ生成の使い道は主に 2 つあると考えられる．まず一つ目は，精度向上のためのデータ拡張である．少数データに対して行ったように，元データが少ない際にはデータを拡張することによって学習データを増やし，汎用性向上に繋げることが可能である．2 つ目は，本物と同等のデータを作り出すことで，データの転用が可能であるという点だ．分析の結果，5000 サンプル以上のデータに関しては，本物データと遜色のないデータを生成することが確認された．つまり，精度が変わらないほどリアルなデータを生成できるため，プライバシーを考慮したデータ分析や効率的なデータ収集が可能になると考えられる．例えば，本物データではなく生成されたデータを用いることで個人情報を保護したり，一部のデータを用いてデータ拡張することで，全てのデータを集計する必要がなくなり効率的な集計が期待される．生成されたデータを用いることでモデルの汎用性向上やプライバシー保護に活用できるだろう．



図–4 活動パターンの精度 (TVAE)



図–5 交通手段と実データの精度 (TVAE)

5. おわりに

今回の分析では，CTGAN,TVAE を用いてアクティビティデータ生成を行い，生成データの活用性の検討を行った．まず，生成されたデータの質に関しては TVAE の方が優れいていることが確認された．KL ダイバージェンスと尤度の指標を比較した際，尤度に関しては CTGAN の方が優れていたものの，KL ダイバージェンスに関しては TVAE の方が優れいている結果となった．しかし，これらの指標のみではテーブルデータ全体の良し悪しは確認できないため，生成されたデータを使用し予測モデルを学習させることで生成データが元データと同様に機能するかどうかを確認した．その結果，PT データにおいては TVAE によるデータセット拡張の方が有効であるという結果となった．しかし，CTGAN によるデータ生成も TVAE には劣るものの生成サンプルを増やしていけば本物データと同等のデータセットを作成することが可能であり，行動データに対しても生成モデルの適用可能性が示唆された．

今度の課題として，PT データに対してのみ検証を行っている点が問題であると考える．本研究では活動パターンと交通手段の 2 つのデータセットを用意し，生成データの機械学習における精度を検証した．しかし，あくまで PT データのみの検証であり，異なるデータセットでの検証は行っていない．そのため，本研究の CTGAN や TVAE はあくまで PT データでの学習になっており，他のデータセットを使用したときには学習結果が異なると予想される．今回の分析では，データセットは 2 種類あるもののあくまで PT データであるため今回のような結果になった可能性も考えられる．今後は異なるデー

タセットでの生成も行う必要があると考えられる．

参考文献

1) Marti, G.: Corrgan: Sampling realistic financial correlation matrices using generative adversarial networks, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8459–8463, IEEE, 2020.
2) Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y.: Data synthesis based on generative adversarial networks, *arXiv preprint arXiv:1806.03384*, 2018.
3) 平湯和也, 河野英昭, 折居英章, and 辻康弘: Gan を用いた擬似データ生成による血中薬物濃度推定, **バイオメディカル・ファジィ・システム学会大会講演論文集** *31*, pp. 49–52, バイオメディカル・ファジィ・システム学会, 2018.
4) Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J.: Generating multi-label discrete patient records using generative adversarial networks, *Machine learning for healthcare conference*, pp. 286–305, PMLR, 2017.
5) Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C.: Veegan: Reducing mode collapse in gans using implicit variational learning, *Advances in neural information processing systems*, Vol.30, 2017.
6) Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification, *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293, IEEE, 2018.
7) Xu, L. and Veeramachaneni, K.: Synthesizing tabular data using generative adversarial networks, *arXiv preprint arXiv:1811.11264*, 2018.
8) Bai, S., Kolter, J. Z., and Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, *arXiv preprint arXiv:1803.01271*, 2018.
9) Patki, N., Wedge, R., and Veeramachaneni, K.: The synthetic data vault, *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, IEEE, 2016.
10) Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114*, 2013.

(? 受付)

Development of Activity Data Generation Method using CTGAN and TVAE

Yushi ISHIJIMA, Hideki YAGINUMA, Shintaro TERABE, Haruka UNO, Yu SUZUKI

In recent years, there have been many research cases in which machine learning theories such as neural networks are used in traffic behavior analysis. When using these methods, it is desirable that a huge amount of learning data is required and that the quality and quantity of the data are guaranteed. However, the cost of acquiring individual activity and traffic behavior data is still high, and it is difficult to secure sufficient data. In this study, we propose an extension method of PT data using GAN (Generative Adversarial Network) and try to generate a dataset with the metropolitan area as a case study. Specifically, by applying CTGAN incorporating LSTM (Long Short Term Memory) and TVAE utilizing Autoencoding, it became possible to generate data that retains activity information before and after. This is expected to improve the efficiency of model learning and data collection in small samples.